Style■Controllable Speech■Driven Gesture Synthesis Using Normalising Flows

Year: 2020 | Citations: 202 | Authors: Simon Alexanderson, G. Henter, Taras Kucherenko, J. Beskow

Abstract

Automatic synthesis of realistic gestures promises to transform the fields of animation, avatars and communicative agents. In off■line applications, novel tools can alter the role of an animator to that of a director, who provides only high■level input for the desired animation; a learned network then translates these instructions into an appropriate sequence of body poses. In interactive scenarios, systems for generating natural animations on the fly are key to achieving believable and relatable characters. In this paper we address some of the core issues towards these ends. By adapting a deep learning■based motion synthesis method called MoGlow, we propose a new generative model for generating state■of■the■art realistic speech■driven gesticulation. Owing to the probabilistic nature of the approach, our model can produce a battery of different, yet plausible, gestures given the same input speech signal. Just like humans, this gives a rich natural variation of motion. We additionally demonstrate the ability to exert directorial control over the output style, such as gesture level, speed, symmetry and spacial extent. Such control can be leveraged to convey a desired character personality or mood. We achieve all this without any manual annotation of the data. User studies evaluating upper■body gesticulation confirm that the generated motions are natural and well match the input speech. Our method scores above all prior systems and baselines on these measures, and comes close to the ratings of the original recorded motions. We furthermore find that we can accurately control gesticulation styles without unnecessarily compromising perceived naturalness. Finally, we also demonstrate an application of the same method to full■body gesticulation, including the synthesis of stepping motion and stance.